ECO375 - Applied Econometrics

Heather Tan

Sep - Dec 2019

Contents

1 Review of Statistics

1.1 Steps of problem solving

1. give an question
2. set out a model
3. propose the estimator
4. check whether the estimator is good
5. if good, how to do inference?(confidence interval, hypothesis testing)

1.2 Problem solving example

Question: Average income in Canada

1.2.1 Model

a Probability model: X = income and $X \sim N(\mu, \sigma^2)$ while μ is known and σ^2 is unknown.

b Sample is $\{x_1, x_2, \cdots, x_n\}$ and assume a random sample iid(identically independently distribution)
Identically: same population
Independently: known about first guy provide no information about the next one.

1.2.2 Propose an Estimator

Definition 1.1. A statistics is a function of the data.

Definition 1.2. An estimator is a statistic that is used to guess the parameter of interest

In this question, parameter of interest is μ
Proposed estimator: sample average: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$

1.2.3 Is this a good estimator?

To answer this we need to know the sampling distribution of the estimator

How do we find the sample distribution?
$E[\bar{X_n}] = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}E[\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}n \cdot \mu = \mu$

$\implies \bar{X_n}$ is an unbiased estimator of μ

$$
\begin{aligned}
Var(\bar{X_n}) &= Var(\frac{1}{n}\sum_{i=1}^{n} X_i) \\
&= \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i), iid \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\
&= \frac{1}{n^2}\cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}
\end{aligned}
$$

An estimator is called **consistent** when its sampling distribution becomes more and more concentrated around the parameter of interest as the sample size increase.

Note that $\bar{X_n}$ is a consistent estimator of μ: $Var(\bar{X_n} = \frac{\sigma^2}{n} \to 0, as\ n \to \infty$

What about $\bar{X_n} \sim$? Fact:

$$
\begin{aligned}
Y_1 &\sim N(\mu_1, \sigma_1^2) \\
Y_2 &\sim N(\mu_2, \sigma_2^2) \\
\implies Y_1 + Y_2 &\sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2cov(y_1, y_2)) \\
\bar{X_n} &= \frac{1}{n}\sum_{i=1}^{n} X_i \sum N(\mu, \frac{\sigma^2}{n})
\end{aligned}
$$

1.2.4 how to do inference

CI: where the parameter is likely to lie in relation to the estimate

Fact:

$$
\begin{aligned}
E[Y] = 0, Var(Y) = \sigma^2 &\implies Z = \frac{y-\mu}{\sigma}, E[Z] = 0, Var(Z) = 1 \\
Z &= \frac{\bar{X_n} - E[\bar{X_n}]}{\sqrt{Var(\bar{X_n})}} = \frac{\bar{X_n} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)
\end{aligned}
$$

$$
\begin{aligned}
1 - \alpha &= P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) \\
&= P(-Z_{\alpha/2} \le \frac{\bar{X_n} - \mu}{\sqrt{\sigma^2/n}} \le Z_{\alpha/2}) \\
&= P(\bar{X_n} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X_n} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \\
(1-\alpha)\%CI &= [\bar{X_n} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X_n} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]
\end{aligned}
$$

2 Simple Regression - Model, Estimate OLS, Properties of OLS

2.1 Econometric model

(Y,X,U) are random variables with joint distribution: $y = g(x, u)$

- Y: dependent variable
- X: explanatory variable
- U: unobserved variable

Facts:

- Summations:
  - $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
  - $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$
- Law os iterated expectations: $E(y) = E[E(Y|X)]$

Want to knwo: $\frac{\partial y}{\partial x} = \frac{\partial g}{\partial x}|_v$

2.1.1 SLR Assumption 1: linear in parameters

$$y = \beta_0 + \beta_1 x + u$$
$$\frac{\partial y}{\partial x} = \beta_1$$

parameters (β_0, β_1)

E.g

$$y = \beta_0 + \beta_1 x^2 + u \implies \frac{\partial y}{\partial x} = 2\beta_1 x$$
$$log(y) = \beta_0 + \beta_1 log(x) + u \implies \frac{\partial log(y)}{\partial log(x)} = \beta_1 \approx \frac{\Delta y\%}{\Delta x\%}$$

2.1.2 SLR Assumption 2: Zero Conditional Mean

1. E[U—X] = E[U]
2. E[U]=0

E.g.1 y = wage, x = training program, u = abilities

If training is assigned randomly

$\implies$ X and U are fully independent

$\implies$ A2.1 implies

E.g.2 y = wage, x = education, u = abilities

$E[U|x=0] \neq E[U|X=1]$

$\implies$ A2.1 is violeted

Implication of A1 and A2

$$\begin{aligned}
E[y|x] &\overset{A_1}{=} E[\beta_0 + \beta_1 x + u|x] \\
&= \beta_0 + \beta_1 E[x] + E[u|x] \\
&= \beta_0 + \beta_1 x + E[u] \\
&= \beta_0 + \beta_1 x \\
&\implies E[y|x] = \beta_0 + \beta_1 x
\end{aligned}$$

conditional expectation is called regression funtion

2.1.3 SLR Assumption 3: Random Sample

$(x_1, y_1), (x_2, y_2), \cdots, x_n, y_n$ i.i.d

2.1.4 SLR Assumtion 4: No Perfect Collinearity

$\{x_1, x_2, \cdots, x_n\}$ are not all the same(sample variation)

2.2 Estimate OLS

Idea: choose your estimator of β_0 and β_1 to minimize dthe su of the square of the errors.

$$\begin{aligned}
min\ Q(b_0, b_1) &= min\ \sum_{i=1}^{n}(y_1 - b_0 - b_1 x)^2 \\
\frac{\partial Q}{\partial b_0} &= -\sum_{i}^{n} 2(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (i) \\
\frac{\partial Q}{\partial b_1} &= -\sum_{i}^{n} 2(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \\
\text{From(i) } \frac{\sum y_i}{N} &- \frac{\sum \hat{\beta}_0}{N} - \frac{\sum \hat{\beta}_1 x_i}{N} = 0 \\
\bar{y} &- \frac{N \cdot \hat{\beta}_0}{N} - \hat{\beta}_1 \bar{x} = 0 \\
&\implies \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (ii) \\
\text{From(i)-(ii)} &\sum_i (y_i - (\bar{y} - \hat{\beta}_1 x) - \hat{\beta}_1 x_i)x_i = 0 \\
&\implies \sum_i (y_i - \bar{y} - (\hat{\beta}_1(\bar{x} - x_i))x_i = 0 \\
&\implies \sum_i (y_i - \bar{y})x_i - \hat{\beta}_1 \sum (x_i - \bar{x})x_i = 0 \\
&\implies \hat{\beta}_1 = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})} = \frac{\sum (x_i - \bar{x})y_i}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

2.3 Properties of OLS - Is OLS a good estimator?

2.3.1 Expected value of $\hat{\beta}_1$

conditional on $x_1, x_2, \cdots, x_n$

$$\begin{aligned}
E[\hat{\beta}_1 | x_1, x_2, \cdots, x_n] &= E[\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} | x_1, x_2, \cdots, x_n] \\
&= \beta_1 + E[\frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} | x_1, x_2, \cdots, x_n] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sum_{i=1}^{n} E[(x_i - \bar{x})u_i | x_1, x_2, \cdots, x_n] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sum_{i=1}^{n} (x_i - \bar{x}) E[u_i | x_1, x_2, \cdots, x_n] \\
&\implies E[\hat{\beta}_1 | x_1, x_2, \cdots, x_n]] = \beta_1
\end{aligned}$$

Law of iterated expectations: $E[y] = E[E[y|x]]$

i.e average the average given by groups $\implies E[\hat{\beta}_1] = E[E[\hat{\beta}_1 | x_1, x_2, \cdots, x_n]] = E[\beta_1]$ by LIE

2.3.2 Variance of $\hat{\beta}_1$

$$\begin{aligned}
Var(\hat{\beta}_1) &= Var(\beta_1 + \frac{\sum(x_i-\bar{x})u_i}{\sum(x_i-\bar{x})^2}) = Var(\frac{\sum(x_i-\bar{x})u_i}{\sum(x_i-\bar{x})^2}) \\
&= \frac{1}{[\sum(x_i-\bar{x})^2]^2} \cdot Var(\sum(x_i-\bar{x})u_i) \\
Var(\sum(x_i-\bar{x})u_i) &= \sum_{i=1}^{n} Var(x_i-\bar{x})u_i) \\
&= \sum_{i=1}^{n}(x_i-\bar{x})^2 Var(u_i|x_1, x_2, \cdots, x_n) \\
&= \sum_{i=1}^{n}(x_i-\bar{x})^2 Var(u_i|x_i) \\
&\overset{A5}{=} \sigma_u^2 \sum_{i=1}^{n}(x_i-\bar{x}) \\
Var(\hat{\beta}_1) &= \frac{1}{[\sum(x_i-\bar{x})^2]^2} \cdot \sigma_u^2 \sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{\sigma_u^2}{\sum(x_i-\bar{x})^2}
\end{aligned}$$

2.4 SLR Assumption 5: Homoscedasticity

$Varu|x = \sigma_u^2 \implies Var(y|x) = Var(\beta_0 + \beta_1 x + u|x) = Var(u|x) = \sigma_u^2$

2.5 Gauss-Markov Theorem

Under assumption A1-A5, OLS(Ordinary Least Square) is BLUE(Best Linear Unbiased Estimator)

2.6 Standard Error

let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

and $\hat{u}_i = y_i - \hat{y}_i$

Define $\hat{\sigma}_u^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{u}_i^2)$

Standard Error: $se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}_u^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

(without hat on σ_u^2 would be sd of $\hat{\beta}_1$)

2.7 Algebraic Properties of OLS

- $\sum_{i=1}^{n}\hat{u}_i = 0$
- $\sum_{i=1}^{n}\hat{u}_i x_i = 0$
- $R^2 = 1 - \frac{SSR}{SST} \in [0, 1]$ where $SSR = \sum_{i=1}^{n}(\hat{u}_i^2), SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$

3 Multi-Linear Regression

3.1 MLR Assumption1-Assumption4

- MLR Assumption 1: $y = \beta_0 + \beta_1 x_1 + \cdots, +\beta_k x_k + u = [1, x_1, x_2, \cdots, x_k] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_k \end{bmatrix} + u \implies y = X\beta + u$
- MLR Assumption 2: $E[u|x_1, \cdots, x_n] = 0$
- MLR Assumption 3: IID Data $\{yi, x_{1i}, x_{2i}, \cdots, x_{ki} \quad i = 1, \cdots, N\}$
- MLR Assumption 4: No perfect Collinearity - There is no exact linear relationship among the explanatory variables

E.g of perfect collinearity:

y = share of votes for A

x_1 = Advertisement Expenditure for A x_2 = Adv. Expenditure for B x_3 = Total Adv. Expenditure

$$y \uparrow = \beta_0 + \beta_1 x_1 \uparrow + \beta_2 x_2 + \beta_3 x_3 + u$$
$$x_3 = x_1 \uparrow + x_2 \downarrow, x_3 \text{ not } \perp x_1$$
$$y = \beta_0 + (\beta_1 + \beta_3) x_1 + (\beta_2 + \beta_3) x_2 + u$$

3.2 Estimator OLS

$\overset{min}{(b\_.b_1, \cdots, b_k)} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2 \implies$ OLS is the solution fo this system of equations.

3.3 Algebraic Properties of OLS:

- $R^2 \uparrow$ when we add more explanatory variables
- partially out(get the effect of x_1 out of other x)
  - 1st: regress x_1 on all other $x_2, x_3, \cdots, x_k$ and get $x_1 = \alpha_1 + \alpha_2 x_{2i} + \cdots + \alpha_k x_{ki} + \Omega$.
    Get residuals $\hat{\Omega} = x_{1i} - (\alpha_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k)$
  - 2nd: regress y_i on $\hat{\Omega}_i \implies y_i = r_o + r_1 \hat{\Omega}_i + v_i$. Thus $\hat{r}_1 = \frac{\sum_{i=1}^{n} (\hat{\Omega}_i) y_i}{\sum_{i=1}^{n} \Omega^2} = \hat{\beta}_1$
    Interpretation of $\hat{r}_1$: The variation of x_1 that cannot be explained by other explanatory variables, which is the part of x_1 that is uncorrelated with $x_2, x_3, \cdots, x_k$

3.4 Statistical properties of OLS

Theorem 3.1. *Under A1-A4, OLS is unbiased -* $E[\hat{\beta}_j] = \beta_j, j = 0, 1, \cdots, k$

3.4.1 Omitted Variable Bias

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\ E[u|x_1, x_2] = 0 \end{cases}$$

But suppose you ignore x_2 and instead consider $y = \beta_0 + \beta_1 x_1 + u$

E.g: Y = incidence of cancer, x_1 = coffee, x_2 = smoking

$$\begin{cases} x_2 = \alpha_0 + \alpha_1 x_1 + v \\ E[v|x_1] = 0 \end{cases}$$

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(\alpha_0 + \alpha_1 x_1 + v) + u \\ y &= (\beta_0 + \beta_2\alpha_0) + (\beta_1 + \beta_2\alpha_1)x_1 + \beta_2 v + u \\ \implies y &= \delta_0 + \delta_1 x_1 + \Sigma \end{aligned}$$

If $E[\Sigma|x_1] = 0 \implies$ OLS $\hat{\delta_1}$ for δ_1 is unbiased.

$E[\hat{\delta}_1] = \delta_1 = \beta_1 + \beta_2 \cdot \alpha_1 > \beta_1$ i.e δ_1 is an biased estimator of β_1.

3.4.2 MLR Assumption 5: Homoscedasticity

$$Var(u|x_1, x_2, \cdots, x_k) = \sigma_u^2$$

Implication: $Var(y|x) = \sigma_u^2$

Theorem 3.2. *Under A1-A5:* $Var(\hat{\beta}_j) = \frac{\sigma_u^2}{[\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2][1-R_j^2]}, for\ j = 1, 2, \cdots, k$

When R_j^2 is the R^2 of the regression of X_j on all other x's.

3.5 Decision on Adding Explanatory Variables

Supposed we have $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$, should we divide u to $u = \beta_{k+1}x_{k+1} + v$ and get $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1}x_{k+1} + v$?

Possibilities:

1. if $\beta_{k+1} = 0$, then we should **NOT INCLUDE** x_{k+1}
2. if $\beta_{k+1} \neq 0$ and x_{k+1} is uncorrelated with all other x's
    - Bias? There is no omitted variable bias problem here
    - Var? Note that $R_j^2 = 0 \wedge \sigma_u^2 \downarrow \implies Var(\hat{\beta}_1) \downarrow \implies$ not solving bias problem, but decrease the variance here.
    - **INCLUDE** x_{k+1}
3. if $\beta_{k+1} \neq 0$and x_{k+1} is correlated with other x's

- Bias? Excluding x_{k+1} leads to omitted variable bias.
- Var? Note that $R_j^2 \uparrow \implies Var(\hat{\beta}_1) \uparrow \ \wedge \ \sigma_u^2 \downarrow \implies Var(\hat{\beta}_1) \downarrow \implies$ Unclear whether $Var(\hat{\beta}_1)$ increase or decrease

3.6 MLR Assumption 6: Normality

$$U \sim N(0, \sigma_u^2)\text{conditional on X}$$

3.6.1 Implication

$$y = X\beta + u$$
$$y|x \sim N(X\beta, \sigma_u^2)$$

3.6.2 Sampling Distribution of OLS

Theorem 3.3. *Under A1-A6:* $\hat{\beta}_j \sim N(\beta_j, \frac{\sigma_u^2}{[\sum_{i=1}^n (x_{ij}-\bar{x}_j)^2][1-R_j^2]}) \implies \frac{\hat{\beta}_j - \beta_j}{\sqrt{Var(\hat{\beta}_j)}} \sim N(0,1)$ *for* $j = 1, 2, \cdots, k$

Note that:

$$\hat{\sigma_u^2} = \frac{1}{n-k-1}\sum_{i=1}^n (\hat{u}_i)^2, \hat{u}_i = y_i - x_i\hat{\beta}$$
$$se(\hat{\beta}_j) = \sqrt{\frac{\sigma_u^2}{[\sum_{i=1}^n (x_{ij}-\bar{x}_j)^2][1-R_j^2]}} = \sqrt{\widehat{Var(\hat{\beta}_j)}}$$
$$\implies \frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{Var(\hat{\beta}_j)}}} = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim \tau(n-k-1)$$

3.6.3 Confidence Interval

$$T = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim \tau(n-k-1)$$
$$1 - \alpha = Pr(-C \leq T \leq C)$$
$$= Pr(\hat{\beta}_j - c \cdot se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c \cdot se(\hat{\beta}_j))$$
$$\implies (1-\alpha)\% CI \ is \ [\hat{\beta}_j - c \cdot se(\hat{\beta}_j), \hat{\beta}_j + c \cdot se(\hat{\beta}_j)]$$

4 T-Test

4.1 Hypothesis Testing

$$H_0 : \beta_j = \beta_j^o$$
$$H_1 : \beta_j \neq \beta_j^o$$

Two types of errors:

1. Reject H0 when H0 is true
2. Reject H1 when H1 is true

Trade off: $\downarrow Pr(Rej\ H_0|H_0) \implies \uparrow Pr(Rej\ H_1|H_1)$

Asymetric: fix the $Pr(Rej\ H_0|H_0)$ at a very small level(Reject H0 when there is stron gevidence to against it)

4.1.1 4 Steps of doing T-test

1. Fix the $Pr(Rej\ H_0|H_0)$ at some level, say α
2. Define a test-statistics: $T = \frac{\hat{\beta_j}-\beta_j}{se(\hat{\beta_j})} \sim \tau(n-k-1)$ if $\beta_j = \beta_j^o$. We have to knwo the distribution of test-stat under H0
3. Define the rejection region: Rejection Region $= \{|T| > c\}$ where c is the critical value

$$\begin{aligned}\alpha = Pr(Rej\ H_0|H_0) &= Pr(|T| > c|H_0)\\ &= 1 - Pr(-C \leq T \leq C|H_0 \text{ True})\\ &\rightarrow \text{ get critical value from the table of t-distribution}\end{aligned}$$

4. check
   - calculate $T = \frac{\hat{\beta_j}-\beta_j}{se(\hat{\beta_j})}$ in the sample
   - compute that with critical value c and decide whether reject H0 or not

4.1.2 P-value

P-value: given the observed value of the test-statistic, what is the **smallest** significane level(α) at which the null would be rejected?

Decrease p-value $\implies$ The **greater** the evidence against H0

5 F-Test

For $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$, the joint test test for: last q coefficients are zero

$H_0 : \beta_{k-q+1} = 0, \beta_{k-q+2} = 0, \cdots, \beta_k = 0$

H_1 : at least one of them is not 0

Idea: Under H_0, we have the restricted model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-q} x_{k-q} + u$

Recall: when exclude X's from a regression $\downarrow R^2 = 1 - \frac{SSR\uparrow}{SST}$. We can build a test-statistic based on by how much SSR increase.

5.0.1 F-test Test Statistic

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(n-k-1)} \sim F(q, n-k-1)$$

6 Conclusion in Model Assumptions

6.1 Data Assumption

A3: IID Data

A4: No perfect linearity

6.2 $E[Y|X] = X\beta$

$$E[Y|X] = X\beta \impliedby \begin{cases} A1 & : Y = X\beta + u \\ A2 & : E[u|X] = 0 \end{cases}$$

6.3 $Var(Y|X) = \sigma_u^2$

$$Var(Y|X) = \sigma_u^2 \impliedby \begin{cases} A1 & : Y = X\beta + u \\ A2 & : E[u|X] = 0 \\ A5 & : Var[u|X] = \sigma_u^2 \end{cases}$$

6.4 $Y|X \sim N(X\beta, \sigma_u^2)$

$$Y|X \sim N(X\beta, \sigma_u^2) \impliedby \begin{cases} A1 & : Y = X\beta + u \\ A2 & : E[u|X] = 0 \\ A5 & : Var[u|X] = \sigma_u^2 \\ A6 & : u|X \sim N(0, \sigma_u^2) \end{cases}$$

6.5 OLS unbiased

$$\text{OLS unbiased} \impliedby \begin{cases} A1 & : Y = X\beta + u \\ A2 & : E[u|X] = 0 \\ A3 & : \text{IID Data} \\ A4 & : \text{No perfect linearity} \end{cases}$$

6.6 OLS is BLUE, $\hat{\beta}_j \sim N(\beta_j, \frac{\sigma_u^2}{[\sum(x_{ij}-\bar{x_j})^2](1-R_j^2)})$

$$\text{OLS is BLUE, } \hat{\beta}_j \sim N(\beta_j, \frac{\sigma_u^2}{[\sum(x_{ij}-\bar{x_j})^2](1-R_j^2)}) \impliedby \begin{cases} A1 & : Y = X\beta + u \\ A2 & : E[u|X] = 0 \\ A3 & : \text{IID Data} \\ A4 & : \text{No perfect linearity} \\ A5 & : Var[u|X] = \sigma_u^2 \end{cases}$$

7 STATA Output

Source	ss	df	MS
Model	SSE	k	SSE/k
Residual	SSR	n-k-1	SSR/n-k-1 = σ_u^2
Total	SST = SSE=SSR = $\sum(y_i - \bar{y})$	n-1	SST/n-1

8 Asymptotics

Why care?

Since $x \sim N(\mu, \sigma_u^2)$ and IID Data $\implies \bar{x} = N(\mu, \sigma_u^2), \bar{x} = \frac{1}{n}\sum x_i$.

When $x \not\sim$ Normal, $\implies \bar{x} \not\sim N \implies T = \frac{\bar{x}-\mu}{\sqrt{Var(\bar{x})}} \not\sim N(0,1) \wedge T = \frac{\bar{x}-\mu}{\sqrt{\hat{\sigma^2}/n}} \not\sim \tau(n-1) \implies$ we cannot do CI and T or F testing.

Solution: Use asymptotic theory(or "large samples")

Same problem/Solution for linear regression models: What if A6 unsatisfied that $u|x \not\sim N$?

8.1 Consistency

Definition 8.1. $\hat{\theta}_n$ is a consistent estimator of θ if for every $\epsilon > 0, Pr(|\hat{\theta}_n - \theta| < \epsilon \to 1$ when $n \to \infty)$

Notation: $\hat{\theta}_n \xrightarrow{p} \theta_n$

8.1.1 Law of large number

Let $x_1, x_2, \cdots, x_n$ be iid with mean $\mu = E[x]$

Then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n x_i \xrightarrow{p} E[x] = \mu$ Intuition: $Var(\bar{x}) = \frac{\sigma^2}{n} \to 0$ as $n \to \infty$

8.1.2 Properties

Let $\hat{\theta}_n \xrightarrow{p} \theta \wedge \hat{\alpha}_n \xrightarrow{p} \alpha$

- $\hat{\theta_n} + \hat{\alpha_n} \xrightarrow{p} \theta + \alpha$
- $\hat{\theta_n} \cdot \hat{\alpha_n} \xrightarrow{p} \theta \cdot \alpha$
- $\hat{\theta_n}/\hat{\alpha_n} \xrightarrow{p} \theta/\alpha$ provided $\alpha \neq 0$
- $g(\hat{\theta_n}) \xrightarrow{p} g(\theta)$ provided that $g(\cdot)$ is a continuous function

8.2 Consistency of $\hat{\beta}_j$ in Regression Model

Theorem 8.1. *Under A1-A4, OLS is consistent. i.e $\hat{\beta}_j \xrightarrow{p} \beta_j$ for j=1,2,....,k*

Proof. Since $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i-\bar{x})y_i}{\sum_{i=1}^n (x_i-\bar{x})^2} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})u_i}{\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2}$
we rewrite the **numerator** as

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})u_i &= \frac{1}{n}\sum_{i=1}^n (x_i - E(x_i) + E(x_i) - \bar{x})u_i \\
&= \frac{1}{n}\sum_{i=1}^n (x_i - E(x_i))u_i + \frac{1}{n}\sum_{i=1}^n (E(x_i) - \bar{x})u_i \\
&= \frac{1}{n}\sum_{i=1}^n (x_i - E(x_i))u_i + (E(x_i) - \bar{x})\frac{1}{n}\sum_{i=1}^n u_i
\end{aligned}$$

by A3 iid Data, $\bar{X} = \frac{1}{n}\sum_{i=1}^n X \stackrel{p}{=} EX$

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})u_i &= \frac{1}{n}\sum_{i=1}^n (x_i - E(x_i))u_i + (E(x_i) - \bar{x})\frac{1}{n}\sum_{i=1}^n u_i \\
&\xrightarrow{p} E[(x-E(x))u] + (E(x) - \bar{x})\frac{1}{n}E(u) \\
&= E[(x-E(x))(u-E(u))] + 0 \cdot \frac{1}{n}E(u) \\
&= cov(x,u)
\end{aligned}$$

and the **denominator** can be rewrite as

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2 &= \frac{1}{n}\sum_{i=1}^n [x_i^2 + \bar{x}^2 - 2x_i\bar{x}] \\
&= \frac{1}{n}\sum_{i=1}^n x_i^2 - (\bar{x})^2
\end{aligned}$$

by A3 iid Data, $\bar{X} = \frac{1}{n}\sum_{i=1}^n X \stackrel{p}{=} EX$

$$\frac{1}{n}\sum_{i=1}^n x_i^2 \xrightarrow{p} E[x^2]$$

$$\frac{1}{n}\sum_{i=1}^n (\bar{x})^2 \xrightarrow{p} [E(x)]^2$$

$$\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2 = E[x^2] - [E(x)]^2 = Var(x)$$

In conclusion, under the A3 iid Data assumption, we can imply that

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \xrightarrow{p} \beta_1 + \frac{cov(x,u)}{Var(x)}$$

□

8.3 Asymptotic Distribution

Definition 8.2. Let $\{z_1, z_2, \cdots, z_n, \cdots\}$ be a sequence of random variables s.t for all Z $Pr(Z_n \leq z) = F_{Z_n}(z) \to Fz(z) = Pr(Z \leq z)$ as $n \to \infty$, we say F_z is the asymptotic distribution of Z_n

Notation: $Z_n \overset{a}{\sim} Z$ or $Z_n \xrightarrow{d} Z$
If $Z \sim N(0,1)$ then we denote $Z_n \overset{a}{\sim} N(0,1)$

8.3.1 Central limit theorem

let $x_1, x_2, \cdots, x_n$ be a random sample with mean μ and variance $\sigma^2 \implies \frac{\bar{x}-\mu}{\sqrt{\sigma^2/n}}$

Theorem 8.2. *Under A1-A5*

- $\frac{\hat{\beta}_j - \beta_j}{\sqrt{Var(\hat{\beta}_j)}} \overset{a}{\sim} N(0,1)$
- $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \overset{a}{\sim} N(0,1)$
- *usual CI,T-test, F-tests are asymptotically valid.*

9 Heteroskedasticity(Assumption 5 unsatisfied)

$Var(u|x) \neq \sigma_u^2 \implies$

- from A1-A5 cannot imply the formula: $Var(\hat{\beta}_j) = \frac{\sigma_u^2}{[\sum(x_{ij}-\bar{x}_j)^2](1-R_j^2)}$
- OLS is not guaranteed to be BLUE

Note that if A5 is wrong but you use the T-stat to construct test and CI, then: The test and CI are invalid.

9.1 Het-Robust SE

$$\begin{aligned}
\hat{\beta_1} &= \beta_1 + \frac{\sum(x_i - \bar{x})u_i}{\sum(x_i - \bar{x})^2} \\
Var(\hat{\beta_1}) &= \frac{1}{[\sum(x_i - \bar{x})^2]^2} \cdot Var(\sum(x_i - \bar{x})u_i) \\
&= \frac{1}{[\sum(x_i - \bar{x})^2]^2} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot Var(u_i|x_i)
\end{aligned}$$

Take $Var(u_i|x_i) = \sigma_i^2$

Problem: we would have to estimate all β's and all σ_i^2, Halbut White noticed that

$$\begin{aligned} Var(u|x) &= E[u^2|x] - (E[u|x])^2 \\ &= E[u^2|x] \\ \widehat{Var(\hat{\beta}_1)} &= \frac{1}{[\sum(x_i-\bar{x})^2]^2} \cdot \sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot (\hat{u_1})^2 \text{ when } \hat{u}_i = y - x_i\hat{\beta} \end{aligned}$$

It turns out that $\widehat{Var(\hat{\beta}_1)}$ is a consistent estimator of $Var(\hat{\beta}_j) \implies T = \frac{\hat{\beta}_j-\beta_j}{\widehat{Var\hat{\beta}_1}} \overset{a}{\sim} N(0,1)$

9.2 Generalized Least Squares(GLS)

$$Var(u|x) = \sigma_u^2 \cdot h(x)$$

h(x) has to be positive and known

$$\begin{aligned} &\sigma_u^2 h(x) = Var(u|x) = E[u^2|x] - E[u|x]^2 = E[u^2|x] - 0 \\ &\implies \sigma_u^2 = \frac{E[u^2|x]}{h(x)} = E[\frac{u^2}{h(x)}|x] = E[\frac{u}{\sqrt{h(x)}}|x] \\ &\qquad = E[(u^*)^2|x] = Var(u^*|x) \\ y \cdot \frac{1}{\sqrt{h(x)}} &= \beta_0 \cdot \frac{1}{\sqrt{h(x)}} + \beta_1 x_1 \cdot \frac{1}{\sqrt{h(x)}} + \cdots + \beta_k x_k \cdot \frac{1}{\sqrt{h(x)}} + u \cdot \frac{1}{\sqrt{h(x)}} \\ &\implies y^* = \beta_0 x_0^* + \beta_1 x_1^* + \cdots + \beta_k x_k^* + u^* \end{aligned}$$

while $x_0^* = \frac{1}{\sqrt{h(x)}}$ and $Var(u^*|x) = \sigma_u^2$

OLS applied to the transformed model is BLUE.

9.3 Testing Heteroskadasticity

$$\begin{aligned} &H_0 : Var(u|x) = \sigma_u^2 \iff E[u^2|x] = \sigma_u^2 \\ &H_1 : Var(u|x) \neq \sigma_u^2 \end{aligned}$$

STEPS:

- Regress Y on X and get $\hat{u}_i = y_i - x_i\hat{\beta}$
- Estimate the regression i.e to see whether the residual is correlated with explanatory variales:$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + (\text{Interacting Terms}) + v$
- F-test: $H_0 : \delta_0 = \delta_1 = \cdots = \delta_k = 0$

10 Instrumental Variables

Endogeneity: when $cov(x, u) \neq 0 \implies$ OLS is inconsistent.This can be caused by :

- omitted variable biased
- measurement error
- simultaneity(i.e X can cause Y but Y also cause X)
- sample selection

10.1 Solution of Instrumental Variable in Simple Regression

$y = \beta_0 + \beta_1 x_1 + u, cov(x, u) \neq 0$

Assumption: Z is observed and is s.t

- $cov(z, u) = 0$ i.e Valid Instrmental Variable
- $cov(x, z) \neq 0$ i.e Relevant Instrumental Variable

We call Z an instrumental variable

E.g $y = \beta_0 + \beta_1 x_1 + u$

y = wage/income

x = education

u = unobserved ability

- What if Z = SIN $\implies cov(SIN, u) = 0, cov(SIN, x) = 0 \implies$ violate relevant assumption
- What if Z = IQ $\implies cov(IQ, u) \neq 0, cov(IQ, x) \neq 0 \implies$ violate valid assumption and good proxies are bad IV's
- What if Z = Parents'education $\implies cov(z, u) \neq 0, cov(z, x) \neq 0 \implies$ violate valid assumption
- What if Z = number of siblings $\implies cov(z, u) \neq 0, cov(z, x) \neq 0 \implies$ violate valid assumption
- What if Z = tuition subsidies $\implies cov(z, u) = 0$(depends on how subsidies are allocated), $cov(z, x) \neq 0$

Bottom line: need to discuss

- $cov(x, u) \neq 0$
- $cov(z, u) = 0$
- $cov(x, z) \neq 0$

10.2 Instrumental Variable Estimator

10.2.1 Assumptions

- A1: $y = \beta_0 + \beta_1 x + u$
- A2': $cov(z, u) = 0$
- A4': $cov(z, x) \neq 0$
- A3': iid data $\{(y_i, z_i, x_i), i = 1, 2 \cdots, n\}$

IDEA

Step1:

$$\begin{aligned}
E[y] &= E[\beta_0 + \beta_1 x + u] \\
\implies E[y] &= \beta_0 + \beta_1 E[x] + E[u] \\
\implies E[y] \cdot E[Z] &= \beta_0 E[Z] + \beta_1 E[x]E[Z] + E[u]E[Z]
\end{aligned}$$

Step2:

$$\begin{aligned}
yZ &= \beta_0 Z + \beta_1 xZ + uZ \\
\implies E[yZ] &= \beta_0 E[Z] + \beta_1 E[xZ] + E[uZ]
\end{aligned}$$

Subtract (1) from (2)

$$\begin{aligned}
E[yZ] - E[y]E[Z] &= \beta_0 E[Z] - \beta_0 E[Z] + \beta_1 E[xZ] - \beta_1 E[x]E[Z] + E[uZ] - E[u]E[Z] \\
cov(y, Z) &= \beta_1 cov(x, Z) + cov(u, Z) \\
\implies \beta_1 &= \frac{cov(y, Z)}{cov(x, Z)} - \frac{cov(u, Z)}{cov(x, Z)} \\
\implies \beta_1 &= \frac{cov(y, Z)}{cov(x, Z)}
\end{aligned}$$

10.3 Proposed Estimator

use the sample analogs:

$$\hat{\beta_1}^{IV} = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

Consistency?: Under A1, A2', A3', A4'

$$\hat{\beta_1}^{IV} \xrightarrow{p} \frac{cov(z, y)}{cov(z, x)} = \beta_1$$

$cov(z,u) \neq 0 \implies$ IV is inconsistent

Bias?:

$$\begin{aligned}
E[\hat{\beta_1}^{IV}] &= E[\frac{\frac{1}{n}\sum_{i=1}^n (z_i-\bar{z})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^n (z_i-\bar{z})(x_i-\bar{x})}] \\
&\neq \frac{E[\frac{1}{n}\sum_{i=1}^n (z_i-\bar{z})(y_i-\bar{y})]}{E[\frac{1}{n}\sum_{i=1}^n (z_i-\bar{z})(x_i-\bar{x})]} = \frac{cov(z,y)}{cov(z,x)} = \beta_1 \\
&\implies \hat{\beta_1}^{IV} isbiased
\end{aligned}$$

10.4 Variance of $\hat{\beta_1}^{IV}$

A5': $Var(u|z) = \sigma_u^2$

$$Var(\hat{\beta_1}^{IV}) = \frac{\sigma_u^2}{n\sigma_x^2\rho_{xz}^2}$$
$$\widehat{Var(\hat{\beta_1}^{IV})} = \frac{\hat{\sigma_u^2}}{\sum_{i=1}^n (x_i-\bar{x})^2 R_{xz}^2}$$

while R_{xz}^2 is the R^2 obtained by regressing x on z.

Under A1, A2'-A5':

$$T = \frac{\hat{\beta_1}^{IV} - \beta_1}{\sqrt{\widehat{Var(\hat{\beta_1}^{IV})}}} \overset{a}{\sim} N(0,1)$$

$\implies$ usual T-test, F-test, CI are asymptotically valid.

There exists Het-robust se for $\hat{\beta_1}^{IV}$

10.5 Compare OLS and IV

Case 1:

If $cov(x,u) = 0$ and $cov(z,u) = 0$

Then $\begin{cases} \hat{\beta}_{OLS} \xrightarrow{p} \hat{\beta} \wedge \hat{\beta}_{IV} \xrightarrow{p} \beta \\ \widehat{Var(\hat{\beta}_{1OLS})} = \frac{\hat{\sigma_u^2}}{\sum_{i=1}^n (x_i-\bar{x})^2} \leq \frac{\hat{\sigma_u^2}}{\sum_{i=1}^n (x_i-\bar{x})^2 R_{xz}^2} = \widehat{Var(\hat{\beta_1}^{IV})} \text{ Homos Case} \\ \implies \text{use OLS!} \end{cases}$

Case 2:

If $cov(x,u) \neq 0$ and $cov(z,u) = 0$

Then $\begin{cases} \hat{\beta}_{OLS} \overset{p}{\nrightarrow} \hat{\beta} \wedge \hat{\beta}_{IV} \xrightarrow{p} \beta \\ \widehat{Var(\hat{\beta}_{1OLS})} = \frac{\hat{\sigma_u^2}}{\sum_{i=1}^n (x_i-\bar{x})^2} \leq \frac{\hat{\sigma_u^2}}{\sum_{i=1}^n (x_i-\bar{x})^2 R_{xz}^2} = \widehat{Var(\hat{\beta_1}^{IV})} \\ \implies \text{use IV!} \end{cases}$

Case 3:

If $cov(x,u) \neq 0$ and $cov(z,u) \neq 0$

$$\hat{\beta}_{OLS} \xrightarrow{p} \beta + \frac{cov(x,u)}{Var(x)}$$
$$\hat{\beta}_{IV} \xrightarrow{p} \beta + \frac{cov(z,u)}{cov(z,x)}$$

Even if $|cov(z,u)| < |cov(x,u)|$, if $cov(z,x)$ is very small $\implies |\frac{cov(x,u)}{Var(x)}| < |\frac{cov(z,u)}{cov(z,x)}|$ in which OLS would be better.

But in general, it is unclear which one is better.

10.6 Weak Instruments

when $cov(z,x)$ is small, we say the instrument z is weak.

10.6.1 problems from weak IVs

- $\widehat{Var(\hat{\beta}_1^{IV})} = \frac{\hat{\sigma_u^2}}{\sum_{i=1}^n (x_i - \bar{x})^2 R_{xz}^2}, \quad R_{xz}^2 \approx 0 \implies \widehat{Var(\hat{\beta}_1^{IV})} \uparrow$
- $T = \frac{\hat{\beta}_1^{IV} - \beta_1}{\sqrt{\widehat{Var(\hat{\beta}_1^{IV})}}} \overset{a}{\not\sim} N(0,1) \implies$ usuals CI, T,R are invalid
- if $cov(z,u) \neq 0 \implies \hat{\beta}_1^{IV} = \beta + \frac{cov(z,u)}{cov(z,x)}$ is big.

10.6.2 Rule of Thumb

How do we detect weak IV?

Regress X on Z, and get the F-Stat

$$F-stat < 10 \implies \text{Z is a weak IV}$$
$$F-stat \geq 10 \implies \text{Z is a strong IV}$$

10.7 General Case of IV

Structural Equation

$$y = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + \cdots + \beta_k z_{k-1} + u$$

- x is endogenous regression ($cov(x,u) \neq 0$)
- $z_1, z_2, \cdots, z_{k-1}$ are exogenous regression ($cov(z_j,u) = 0, forj = 1, \cdots, k-1$)
- $z_k, z_{k+1}, \cdots, z_q$ are instrumental variables ($cov(z_j,u) = 0, forj = k, \cdots, q$)
- $z = (z_1, \cdots, z_{k-1}, z_k, \cdots, , z_q)$ is exogenous

Reduced from Equation: $(cov(z_j, v) = 0)$i.e write an endogenous variable in terms of exogenous variables.

$$x = \pi_0 + \pi_1 z_i + \pi_2 z_2 + \cdots + \pi_q z_q + v = z\pi + v$$

IDEA: Problem: $cov(x, u) \neq 0$

$$\begin{aligned}
0 \neq cov(x, u) &= cov(z\pi + v, u) \\
&= cov(\pi_0 + \pi_1 x_1 + \cdots + \pi_q z_q + v, u) \\
&= \pi_1 cov(z_1, u) + \pi_2 cov(z_2, u) + \cdots + \pi_q cov(z_q, u) + cov(v, u) \\
&= cov(v, u)
\end{aligned}$$

10.8 2 Stages Least Squares(2SLS)

1st stage: Regress x on all z's $(z_1, \cdots, z_q)$ and get $\hat{x_i} = z_i\hat{\pi}$

2nd stage: Regress y on x_i and all exogenous regressors $z_1, z_2, \cdots, z_{k-1}$(no IV included)

$$\begin{aligned}
y &= \beta_0 + \beta_1 x + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + u \\
&= \beta_0 + \beta_1(z\pi + v) + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + u \\
&= \beta_0 + \beta_1(\pi_0 + \pi_1 x_1 + \cdots + \pi_q z_q + v) + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + u \\
\implies y &= (\beta_0 + \beta_1\pi_0) + (\beta_1\pi_1 + \beta_2)z_1 + \cdots + (\beta_1\pi_{k-1} + \beta_k)z_{k-1} \\
&+ (\beta_1\pi_k)z_k + \cdots + (\beta_1\pi_q)z_q + (\beta_1 v + u) \\
y &= \alpha_0 + \alpha_1 z_1 + \cdots + \alpha_q z_q + \epsilon \\
x &= \pi_0 + \pi_1 z_1 + \cdots + \pi_q z_q + v
\end{aligned}$$

Note:

$$\begin{aligned}
\alpha_1 &= \beta_1\pi_1 + \beta_2 \\
\alpha_2 &= \beta_1\pi_2 + \beta + 3
\end{aligned}$$

For $z_k : \alpha_k = \beta_1\pi_k \implies \beta_1 = \frac{\alpha_k}{\pi_k}$
For $z_{k+1} : \beta_1 = \frac{\alpha_{k+1}}{\pi_{k+1}}$

10.9 Exogenity Test

Want to test $cov(x, u) = 0$

Problem: Cannot use OLS to test this. Recall sample covariance between x_i and $\hat{u_i} = y_i - x_i\hat{\beta}$ is zero by construction $\sum x_i\hat{u_i} = 0$

But can use a valid instrument to test it.

Idea:

$$y = \beta_0 + \beta_1 x + u$$
$$x = \pi_0 + \pi_1 z + v$$
$$cov(z, u) = cov(z, v) = 0$$

Recall: $0 \neq cov(x, u) = cov(v, u)$(because $cov(z, v) = 0$), take $u = \delta v + e$ so that if $cov(x, u) \neq 0, \delta \neq 0 \implies y = \beta_0 + \beta_1 x + \delta v + e$

Steps:

- Regress x on z(OLS) and get $\hat{v_i} = x_i - z_i\hat{\pi}$
- Regress y_i on x_i and $\hat{v_i}$ (OLS)
- Test $H_0 : \delta = 0$ vs $H_1 : \delta \neq 0$ (use a t-test)

Observation:

- called control function
- control function and 2SlS are numerically equivalent
- This depends on $cov(z, u) = 0$

Problem: Cannot use IV estimator to test $cov(z, u) = 0$. In the data, the covariance between z_i and $\hat{u}_i$ is zero by construction $\sum_{i=1}^{n} z_i\hat{u}_i = 0, \hat{u}_i = y_i - x_i\hat{\beta}^{IV}$

10.9.1 Over-Identification Test

$$y = \beta_0 + \beta_1 x + v$$
$$x = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v$$
$$cov(z_1, u) = cov(z_2, u) = cov(z_1, v) = cov(z_2, v) = 0$$

IDEA If you run 2SlS using z_1 only, $\implies \tilde{\beta}_1^{IV}$
If you run 2SLS using z_2 only, $\implies \check{\beta}_1^{IV}$
Thus under the assumption that both z_1 and z_2 are valid IVs. We should have $\tilde{\beta}_1^{IV} \approx \check{\beta}_1^{IV}$
If they are very different, then either z_1 or z_2 or both z_1 and z_2 are invalid.
Key: is to test the distance between, if the difference is to big, then reject.

1. Estimate structural equation using all IV's and get $\hat{u}_i$
2. Regress $\hat{u}_i$ on all exogenous variables (using OLS) and get R^2

3. Under $H_0 : NR^2 \sim \chi_q^2$ where q = number of estimators =. number of endogenous regressors, N = number of observations(number of IV should be greater or equal to the number of endogenous x's. Each endo x's shoudl have at least one related IV)

Bad News:

1. If you have one IV, then you cannot test if cov(z,u) = 0
2. If you have more IV's then endo regressors, you can run the over-ID test. But if you reject H0, then you know something is wrong(either z_1 is invalid or z_2 is invalid or both are invalid) but you don't know which one is invalid(no guidance for what to do next).
3. If you don't reject H0(no evidence that something is wrong), that doesn't guarantee that both IVs are valid (they can both be invalid and dilivers similar estimates $\tilde{\beta}_1^{IV} \approx \check{\beta}_1^{IV}$)